The Navigation and Visualisation of Environmental Audio using Zooming Spectrograms

Michael W. Towsey, Anthony M. Truskinger and Paul Roe
Ecoacoustics Research Centre, Electrical Engineering and Computer Science School
Queensland University of Technology,
Brisbane, Australia
m.towsey;a.truskinger;p.roe@qut.edu.au

***Abstract*—Acoustic recordings play an increasingly important role in monitoring terrestrial and aquatic environments. However, rapid advances in technology make it possible to accumulate thousands of hours of recordings, more than ecologists can ever listen to. Our approach to this big-data challenge is to visualize the content of long-duration audio recordings on multiple scales, from minutes, hours, days to years. The visualization should facilitate navigation and yield ecologically meaningful information prior to listening to the audio. To construct images, we calculate acoustic indices, statistics that describe the distribution of acoustic energy and reflect content of ecological interest. We combine various indices to produce false-color spectrogram images that reveal acoustic content and facilitate navigation. The technical challenge we investigate in this work is how to navigate recordings that are days or even months in duration. We introduce a method of zooming through multiple temporal scales, analogous to Google Maps. However, the "landscape" to be navigated is not geographical and not therefore intrinsically visual, but rather a graphical representation of the underlying audio. We describe solutions to navigating spectrograms that range over three orders of magnitude of temporal scale. We make three sets of observations: 1. We determine that at least ten intermediate scale steps are required to zoom over three orders of magnitude of temporal scale; 2. We determine that three different visual representations are required to cover the range of temporal scales; 3. We present a solution to the problem of maintaining visual continuity when stepping between different visual representations. Finally, we demonstrate the utility of the approach with four case studies.**

Keywords—visualisation of acoustic data, visual analytics, ecological acoustics, multi-scale analysis, zooming interface

I. INTRODUCTION

Acoustic recordings play an increasingly important role in monitoring terrestrial and aquatic environments. In fact, recorded audio contributes to several kinds of ecological investigation concerning biodiversity [1], environmental health [2], threatened species, invasive species [3], and climate change [4]. In the case of terrestrial ecosystems, it is fortunate that three major groups of vocal species, birds, insects, and frogs, are also accepted as important indicators of environmental health [5]. In the last few years, increasing awareness of sound in the environment has spawned a new discipline, *soundscape ecology*, which investigates the temporal and spatial distribution of sound through a landscape as it reflects important ecosystem processes and human activity [6].

Audio recordings are an attractive methodology for large-scale monitoring of the environment because recording devices can be deployed in the field for days or weeks on end, obviating the need for regular field visits by expert ecologists [7]. However, while rapid advances in recording and computing technology make it possible to leave unattended acoustic sensors in exposed locations for weeks, even months, of continuous recording, it is clearly impossible for ecologists to listen to any significant fraction of the audio they collect. As a consequence, much effort has been devoted to automated and semi-automated methods of acoustic analysis, a difficult task because there are no constraints on what is included in environmental recordings (compared for example to speech recognition where transmission noise is carefully controlled). Three tasks in particular have received attention: automated species recognition [8-10]; protocols for sampling from audio [11] and the extraction of acoustic indices as surrogates for biological activity [12-14].

Despite the development of automated techniques, the accumulation of environmental recordings still presents a classical big-data problem – data acquisition is easy but data curation, search, analysis, and visualization present considerable problems. A single 24-hour recording, even when compressed as MP3, is over a gigabyte in size. After seven years of collecting recordings from different sites, our lab is now managing the equivalent of 20 years of audio in 329,000 recordings. At 35 TB of data, this would not be considered a big-data problem if it were text, but long-duration (1-24 hours) audio recordings are opaque and impenetrable. Standard audio software is not designed for this situation.

The challenge addressed in this paper is how to *navigate* and *visualize* the content of long-duration audio recordings on multiple scales, from minutes, hours, days to months. Visualization should facilitate navigation and present meaningful information about acoustic content at each scale *prior* to listening to the underlying audio. This approach is based on the intuition that the brain can integrate large amounts of visual information representing an audio recording more rapidly than it can listen to the audio.

II. Related Work

There are four kinds of technical approach to the visualization of sound: the waveform, the spectrogram (and derivatives such as the cepstrogram), semantic icons, and the chromagram. The waveform is typically not useful for environmental recordings because it is dominated by low frequency wind, rain, and machine noises. However the waveform envelope has been useful for classification of frog calls where the signal-to-noise ratio was high [15]. The chromagram typically has an aesthetic motivation (as used in Windows Media Player) but has also been used to reveal the structure in musical compositions [16]. Symbolic approaches have been used to reveal the structure in musical composition [17], to convey the amplitude and frequency content of simple sounds [18] and to help the hearing impaired by representing sound semantics [19, 20]. In fact, these last two studies combine waveform, spectrogram, and symbolic representations to allow multiple simultaneous visualizations of sound in real-time.

In the ecological context, the spectrogram remains the most useful visual representation because it makes important time-frequency information explicit. The typical spectrogram is a grey-scale image whose pixel columns are spectra (obtained from signal frames) and whose rows are frequency bins. Pixel tone encodes acoustic intensity. At a typical time scale of fifty frames per second, it is intuitively easy to link visual features in the spectrogram to acoustic content. However, the standard spectrogram does not scale to long duration recordings. A 24-hour recording shown as a typical standard spectrogram with a temporal scale of 0.02 s/frame (one frame per pixel column), on a current desktop monitor (35.7 px/cm density) would produce an image 1.2 km wide.

It has been demonstrated that long-duration recordings of the environment can be meaningfully represented at convenient temporal/visual scales by first extracting acoustic indices – statistics that summarize the temporal and spectral distribution of acoustic energy in a segment of recording [21]. There is a growing body of work on the ecological uses of acoustic indices. Some are derived from the audio waveform and others from spectral content. Of particular note are the *acoustic complexity index* (ACI, a measure of the average absolute fractional change in signal amplitude from one frame to the next through a recording) [4, 13] and *entropy* (a measure of the temporal and/or spectral dispersal of acoustic energy) [22]. These indices excite interest because they have been demonstrated to be sensitive to biological sources of acoustic activity, especially birds.

The approach taken in [21] is to combine three indices (calculated at a coarse resolution of one value per frequency bin per minute of audio over a 24-hour recording) and map them to the red-green-blue (RGB) color channels respectively. The result is a false-color, long-duration spectrogram, such as those in Fig. 1, having a temporal scale of 60 s/pixel column (hereafter abbreviated to 60 s/px). We must note the distinction between a false-color and a pseudo-color spectrogram. The latter are frequently produced by signal processing packages and map one-dimensional spectral power values to color (instead of grayscale). In this work, false-color is obtained by mapping three *different* indices to each RGB color channel.

To the extent that the three indices capture orthogonal information, a false-color image of an environmental recording will convey more information (reveal more acoustic structure) than a pseudo-color spectrogram. Indeed, the principle motivation for long-duration false-color spectrograms is to reveal more acoustic structure and provide more acoustic context than is achieved by other color-mappings at the same temporal scale. Another use is to verify data integrity. Acoustic sensors in ecological studies are exposed to all kinds of conditions and, in practice, management and visualization of long-duration audio must accommodate corrupted and noisy data. Typical practice in ecoacoustic studies is to 'weed out' manually unwanted audio prior to analysis [1, 13] but such methods do not scale. In this work, we do not remove audio segments that contain clipping or 'noise' due to wind, rain, aircraft, traffic, and other human activity. In the context of soundscape ecology, such sounds are considered 'signal' and not 'noise'.

Although small-scale spectrograms (60 s/px or lower) can convey meaningful ecological information, there remains a 3000-fold scale gap between this and the standard resolution of ~0.02 s/px. In order to assist navigation through a 24-hour recording, intermediate scales are required. This is the same predicament faced by a person driving across a continent with only a pocket map. A large-scale map covering the continent is impractical and unwieldy but the pocket map is no use in built-up areas. The solution explored in this paper to the problem of navigating long-duration recordings is *spectrogram zooming*.

III. Design Considerations

A. Navigation in Electronic Landscapes

Discussions of navigation through electronic landscapes are helpfully quantified using Fitts' Law [23]. Although originally derived from observations of real world pointing tasks, this law has found an important application in computer-human interfaces. It describes an empirical relationship between the *time* taken to navigate between two locations on a computer screen (for example, using a mouse pointer), the *distance* (D) between the two locations and the *size* (W) of the target location. Because the observed relationship is mathematically similar to Shannon's information theorem, the *index of difficulty* (ID) of a navigation task can be defined as:

$$ID = \log_2(D/W + 1)\,, \qquad (1)$$

where the units of ID are information bits [24]. The maximum ID for a feasible task in a flat landscape (electronic or physical) is close to 10 bits because D cannot exceed 1 m (the reach of an arm) and W cannot be much less than 1 mm. However the ID to navigate a 24-hour recording could reach 13-14 bits (where D = 1.2 km and W = 10-100 mm) which implies the task is not feasible.

Three important empirical observations have been made when a zoom facility is added to navigation tasks in

electronic space: 1. zooming enables the 10-bit ID barrier to be exceeded; 2. zooming becomes advantageous when the task ID reaches 8 bits; and 3. with a zooming facility, tasks remain feasible even when their ID reaches 30 bits [24].

B. The Constraints of Visualising Sound

The mapping of sound to images requires the creation of a 'language' that should be expressive but also concise and easy to learn. There is a tension between these requirements. We believe the spectrogram remains the 'language' of choice for visualizing recordings of the environment because it is well understood by ecologists, it is intuitive and it captures the temporal scale smoothly. A symbolic representation (as coded abstract shapes or icons) will require a compromise between the language expressivity and its learnability. There are also problems with the display of icons when the temporal space they occupy on a monitor is more than that of the event they represent.

The motivation to investigate zooming spectrograms came from the ease with which zooming in Google Maps enables one to navigate images of planet Earth. The spectrogram can be considered a 'landscape' in which only the temporal dimension needs to be scaled. However, spectrograms present a more difficult problem because the landscape to be navigated is not intrinsically visual but rather a graphical representation of underlying audio. In this work, we report our solution to two technical challenges: the first is to find visual representations of audio that are meaningful from small temporal scale to large scale; and second, how to preserve continuity of representation navigating from one scale to the next. To the best of our knowledge, this is the first attempt to assist visualization and navigation of long duration audio-recordings using a zooming facility.

IV. Methods

A. Hardware and Audio Data

The results described in this paper were obtained from five recordings. Four of the recordings were each two-hours long and obtained with battery-powered, weatherproof Song Meter boxes (SM2s from Wildlife Acoustics). They were two-channel, sampled at 22.05 or 44.1 kHz and saved in WAC4 format. The recordings were obtained from locations to be described later in text.

A fifth 24-hour recording was obtained on 13th October 2010 in open eucalyptus woodland on the outskirts of Brisbane city, Australia. The sensor was a custom-built Olympus DM-420 digital recorder housed in a weatherproof case. Two external omnidirectional electret microphones were attached to the recorder which was powered by four D-cell batteries, providing up to 20 days of continuous recording. Data were stored internally in stereo MP3 format (128 kbps, 22050 Hz) on high capacity 32GB Secure Digital memory cards [11]. All the sensor boxes were attached to tree trunks at chest height.

B. Signal Processing

For ease of data processing, all recordings were divided into one-minute segments, mixed to mono and, if necessary, down-sampled to 22050 Hz. FFTs were calculated using a frame size of 512 samples and a Hamming window. The spectrum derived from each frame has 256 frequency bins, spanning 11025 Hz (43.06 Hz per bin). It should be noted that almost all the bioacoustic activity of interest is below 9000 Hz.

For the small-scale (zoomed out) spectrograms (those derived from acoustic indices), frame overlap was not used because it interferes with the calculation of acoustic indices. For the large-scale (zoomed in) spectrogram images, an overlap of 71 samples was used to give a frame step of 0.02s.

We have adopted a linear frequency scale despite it giving more prominence to the high frequency band than is apparent to the human ear. The Mel-scale could be adopted but in our experience, it gives too much prominence to the low-frequency band. Birds, which have been the primary focus of our attention to date, mostly call in the mid-frequency band, so we have retained the linear scale.

C. Acoustic Indices

Five acoustic indices were calculated for each of the 256 frequency bins in each spectrogram derived from a short sub-segment of audio. The sub-segment durations varied from 0.2 s to 60 s. Where the segment length was not an exact number of frames, the entirety of the last frame was taken. For example, 0.2 s of audio spans 8.61 non-overlapping frames and consequently the selection was extended to include nine frames.

For each sub-segment of spectrogram, we calculate five spectral indices, two from the amplitude spectrogram and three from the decibel spectrogram (where decibel values were derived from amplitude values using $dB = 20 \cdot \log_{10} A$).

ACI spectrum: For each frequency bin in the amplitude spectrogram, calculate the average absolute fractional change in spectral amplitude from one frame to the next. That is:

$$ACI_f = \sum_i |a_{fi} - a_{f,i-1})| / \sum_i a_{fi} , \qquad (2)$$

where i is an index $0 \ldots N\text{-}1$ over amplitude values, a_{fi}, in frequency bin f, and N depends on the length of the spectrogram selection. $\sum_i a_{fi}$ is a normalizing factor [13].

ENT spectrum: The 'reversed' entropy spectrum. In this context, entropy is a measure of the *dispersal* of acoustic energy within each frequency bin. For each frequency bin, square the amplitude values (convert to acoustic energy), normalize to unit area and treat the values as a probability mass function (pmf). The temporal entropy of the bin values is:

$$H_f = \sum_i \log_2(pmf_{fi}) / \log_2 N , \qquad (3)$$

where i is an index $0 \ldots N\text{-}1$, in frequency bin f, and N is the number of frames in the spectrogram of the sub-segment [1]. The $\log_2(N)$ term normalises for the value of N. Note that the ENT spectrum was 'reversed' ($ENT_f = 1\text{-}H_f$) to yield an

acoustic energy *concentration* index rather than a *dispersal* index.

BGN spectrum: Background noise for each frequency bin is calculated from a dB spectrogram centered on the current audio selection plus 5 s either side, using an additive noise model as described in [25].

POW spectrum: The dB average of the power values in each frequency bin after subtraction of background noise (the BGN values).

EVN spectrum: A count of the number of acoustic event transitions in each frequency bin of the current noise-reduced spectrogram. A transition occurs when the spectral power (the POW value) crosses the 3 dB threshold from one frame to the next in either direction.

D. Preparation of False-color Spectrograms

Spectrograms were prepared from each of the above indices. At the 0.2 s/spectrum (0.2 s/px) scale, a 24-hour spectrogram contains 432,000 spectral columns and at 60s/spectrum it contains 1440 spectral columns. For each temporal scale from 60 s/px to 0.2 s/px inclusive, false-color images were prepared by mapping three of the above indices to RGB color. In order to produce informative images that utilize the full color range, index values must be normalized as described in [21]. One of the research questions addressed in this paper is which index to map to which color at each scale.

It is worth noting here that we have discontinued recording in MP3 because MP3 compression can generate artifacts in the reconstituted WAV signals. These in turn lead to spurious index values (particularly for ACI and ENT) which become apparent as 'unusual' localized shapes in false-color spectrograms. However, the 24-hour MP3 recording used in this study produced relatively few artifacts and they were not enough to discount the results we present.

E. Preparation of High Resolution Spectrograms

Standard spectrograms were prepared at a temporal scale of 0.02 s/px by setting the frame size = 512 and overlap = 71. Spectrograms at scales small than 0.02 s/px but greater than 0.2 s/px were derived from the 0.02 s/px spectrograms by reduction, i.e. aggregating consecutive values as follows:

$$x_{\text{aggregated}} = x_{\text{mean}} + ((x_{\max} - x_{\text{mean}}).(N\text{-}1)/N_{\max})$$

where x_{mean} is the mean of the values to be aggregated, $x_{\max}$ is the maximum of the values to be aggregated, and N is the number of values to be aggregated. Giving some additional weight to the maximum value prevents aggregation 'washing out' the spectrogram. Following aggregation, BGN and POW spectrograms were prepared from each one minute of recording as described previously. How these indices were mapped to colors at each scale is also a research question addressed in this paper.

F. Tiled Images

We used existing third-party software to view zooming spectrograms in an interactive manner. PanoJS3 is a single JavaScript file solution for viewing high resolution images on a basic HTML webpage [26]. It is designed to view panoramic images that have been cut into smaller tiles and expects aggregate (small-scale) images to be pre-cut for the zoomed-out layers. The software works in all major browsers and comes with inbuilt support for zooming and panning. It can be repurposed to show basic maps, high-resolution document scans, or (in this case) high-resolution images that are not aggregated for small scale views in a typical manner. PanoJS3 is a lightweight, cross-platform, portable, and open source solution for showing the zooming spectrograms presented in this paper. Unlike Google Maps, PanoJS3 does not display interpolated images when transitioning from one scale to the next.

For simplicity, we opted to produce square tiles. The spectrograms had 256 frequency bins and therefore a height of 256 pixels. It was helpful to add a title bar (24 pixels high) and a time scale (20 pixels high), yielding images 300 pixels in height. The resulting 300×300 pixel tiles were named in a basic indexing format used by PanoJS.

V. Implementation

The advantage of false-color spectrograms can be observed by comparing the top spectrogram in Fig. 1 with the gray-scale version (Fig. 1, bottom). A 24-hour recording, even in MP3 format, cannot usually be opened by standard desktop software. However, we achieved this by using the well-known open-source audio-editor, Audacity (http://sourceforge.net/projects/audacity/), on a high-performance computer. The gray-scale spectrogram reveals little acoustic structure apart from broad changes in the level of background 'noise'. By comparison, acoustic structure in the false-color spectrogram is clearly delineated. The morning chorus is visible starting around 0440h. The sounds made by Orthoptera species (grasshoppers, crickets, and katydids) appear as tracks during nighttime hours.

Even at this small temporal scale, a surprising amount of information can be discerned. (An annotated version of Fig. 1 can be seen here: http://www.ecosounds.org/Zoom/1-top. The zooming version of this recording can be seen here: http://www.ecosounds.org/Zoom/zoom). Some bird species can be identified because their calls leave traces in consecutive one-minute spectra. For example, crow calls, consisting of stacked harmonics, can be identified at 1000-1010h (see top Fig. 1, and click link for annotated image). The arrival and departure of other bird species is apparent: the Grey Fantail (*Rhipidura albiscapa*) calling in the 5-7 kHz band during the five hours after dawn; the Yellow-faced Honeyeater (*Lichenostomus chrysops*) calling in the 2-4 kHz band around 0938h; the Olive-backed Oriole (*Oriolus sagittatus*) calling around 1645h; and the Striated Pardalote (*Pardalotus striatus*) calling around 1719h. Despite the coarse scale, 'birders' can learn to recognize the arrival and departure of vocal species that leave such traces. For the majority of vocal species however, detection in a spectrogram requires a temporal scale with finer resolution.

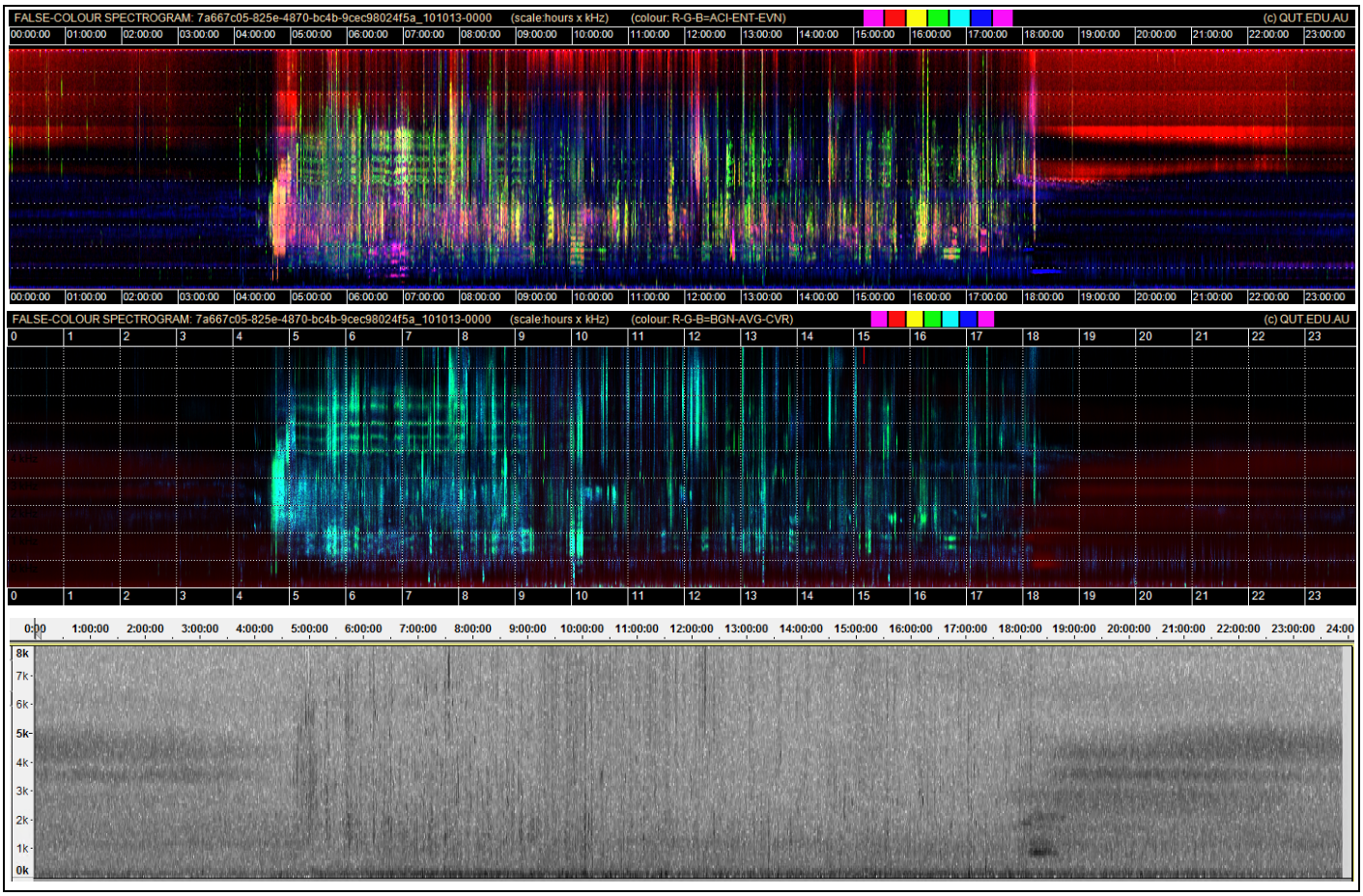

Fig. 1: Three long-duration spectrograms of the same 24-hour recording. **Note 1:** All images have the same temporal scale, midnight to midnight. **Note 2:** Each image has slightly different Hertz scale, see below. **Note 3:** These images view better when enlarged 200%. **Note 4:** See annotated version of top and mid images at http://www.ecosounds.org/Zoom/1-top and http://www.ecosounds.org/Zoom/1-mid. *Top*: A false-color spectrogram that maps the ACI, ENT, and EVN acoustic indices to RGB color (see text). The horizontal gridlines are at 1 kHz intervals, full scale = 11 kHz. For the zooming implementation of this recording, see http://www.ecosounds.org/Zoom/zoom. *Middle*: A false-color spectrogram that maps the BGN, POW, and EVN acoustic indices to RGB color. The horizontal gridlines are at 1 kHz intervals, full scale = 8.8 kHz. *Bottom*: A gray scale spectrogram prepared using the open-source audio-editor, Audacity. Full scale = 8.0 kHz. As determined by a qualified birder, 62 different bird species can be heard in this 24-hour recording and 877 (61%) of the 1440 minutes contain at least one call [11]. A road some 100 meters distant meant that the recording contained muffled traffic noise in addition to the sounds of airplanes, dogs, and moderate wind gusts. There was no rain on the day.

In preparing a set of spectrogram tiles for zoom-assisted navigation, the paramount consideration was to retain an intuitive, visual consistency when stepping between temporal scales. The user should not be surprised or perplexed by a transition and should not have difficulty in keeping focus on a specific acoustic event when zooming in. The underlying computational complexity should be largely 'invisible' because there is little scope in the interface to explain what is happening. Our investigation therefore centered on three questions: 1. How many scale steps are required between the smallest and largest scale? 2. What visual representation is best at each scale? 3. How to preserve continuity between different visual representations at different scales? We describe our observations in turn.

A. The Number of Scale Steps

The determination of an appropriate number of scale steps can once again make use of Fitts' Law and the *index of difficulty*. To quantify the 'gap' between two spectrogram images at different scales, we substituted the *scale ratio* for the D/W term in (1). Note that in this interpretation, the index of difficulty becomes a measure of the potential *user perplexity* in 'jumping the gap' from one scale to the next. The scale gap between images at 60 s/px and 0.02 s/px (the gap we wished to span with a zooming facility) is 11.6 bits, comparable to the navigation ID of 13.6 bits calculated previously.

The Google Maps scale gap between the smallest scale view of Earth and largest scale over Brisbane city is $\log_2(2000\text{ km} / 5\text{ m}) = 18.6$ bits. Google Maps (an interface that has been well honed by user feedback) covers this scale gap in 17 steps, each step on average spanning a gap of 1.1 bits. This suggested that our spectrogram scale gap of 11.6 bits could be covered in 10 or more steps.

However, our experimentation revealed that a greater than two-fold scale step increased user perplexity. As noted above, PanoJS3 does not provide interpolated images in the transition between scale steps. Image interpolation clearly

helps to reduce user perplexity. The scale steps we ultimately implemented are shown in Fig. 2. Only the first scale step exceeded *ID* = 1 bit (scale change > 2, Table 1).

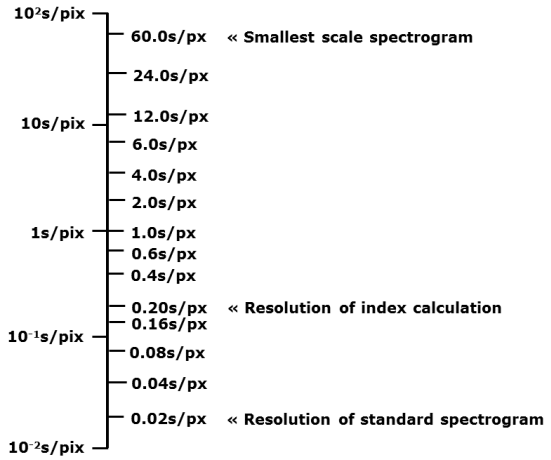

Fig 2: The 14 temporal scales used to span the interval between the smallest spectrogram scale (60 s/px) and the largest scale (0.02 s/px).

B. Visual Representation at Different Scales

Although there may appear to be many ways to map the five spectral indices to RGB color, in fact the possibilities were greatly limited by three constraints. To facilitate the discussion it is helpful to divide the full-scale range into three zones: small-scale (60 s/px to 6 s/px); mid-scale (4 s/px to 0.2 s/px); and large-scale (0.16 s/px to 0.02 s/px).

TABLE 1. Tile information at different scales.

AEE: ACI-ENT-EVN are mapped to RGB respectively; **BPE**: BGN-POW-EVN are mapped to RGB respectively; **BP**: BGN-POW are mapped to red and a color-cube-helix scale, respectively.

Scale (s/px)	Scale step (bits)	300px tile time-span	Aggregation count	Color-map
60	1.5	5 hr	300	AEE
24	1.0	2 hr	120	AEE/BPE 9:1
12	1.0	1 hr	60	AEE/BPE 7:3
6	0.7	30 m	30	AEE/BPE 5:5
4	1.0	20 m	20	AEE/BPE 5:5
2	1.0	10 m	10	AEE/BPE 4:6
1	0.7	5 m	5	AEE/BPE 3:7
0.6	0.7	3 m	3	AEE/BPE 2:8
0.4	1.0	2 m	2	AEE/BPE 1:9
0.2	0.8	1 m	1	BPE
0.16	1.0	48 s	8	BP
0.08	1.0	24 s	4	BP
0.04	1.0	12 s	2	BP
0.02	--	6 s	1	BP

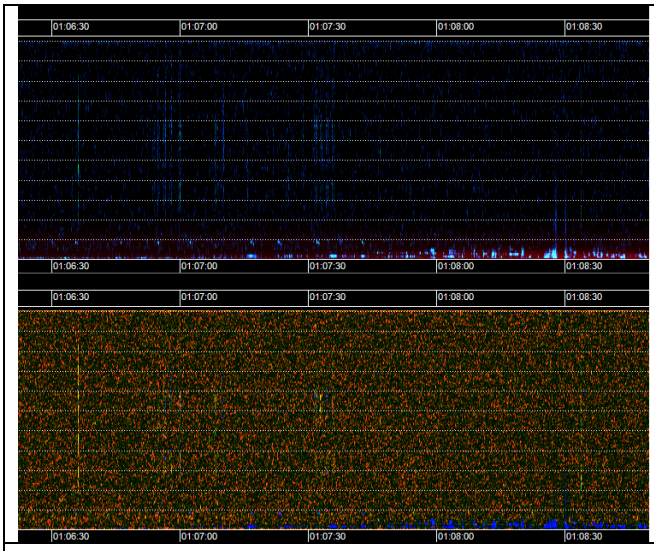

Fig. 3: A comparison of false-color spectrograms at 0.2 s/px scale (approx. two minutes duration) derived from different indices. The top spectrogram is derived from the indices BGN, POW, and EVN. The lower spectrogram is derived from the indices ACI, ENT, and EVN. The variance of the ACI and ENT indices at this temporal scale hide the fine acoustic structure. Horizontal gridlines indicate 1khz intervals.

The first constraint was that the ACI and ENT indices are not well behaved at high resolution (mid and large scales). The acoustic structure in mid and large-scale spectrograms derived from these indices becomes lost in noise (Figure 3, bottom). By contrast, the BGN, POW, and EVN indices continue to reveal acoustic structure at the 0.2s/px scale (Fig. 3, top).

The second constraint was the desirability of combining indices having minimum correlation. It can be observed in Figure 1 that the ACI-ENT-EVN spectrogram at 60s/px reveals more acoustic structure than the BGN-POW-EVN spectrogram because the POW and EVN indices are more highly correlated than any other pair of available indices.

Calculating indices other than BGN and POW for audio segments shorter than 0.2 s was not attempted because a 0.2 s segment only spans 8.6 frames. Only the BGN and POW indices are well behaved at this scale. Consequently, the spectrograms we prepared at scales 0.16, 0.8, 0.04 and 0.02 s/px displayed only BGN and POW, with BGN mapped to red and POW mapped to the *cube-helix* color scale [27]. The commonly used rainbow color map is problematic because there can be large differences in the eye's perception of adjacent map-colors leading to a misrepresentation of the underlying data [28]. Our implementation of the cube-helix color map starts with black (zero acoustic intensity), ends with white (maximum acoustic intensity) and passes through colors for intermediate acoustic values. It preserves a continuous increase in perceived color intensity and thereby avoids the perceptual difficulties posed by most other color maps. The cube-helix map has an additional advantage that it prints as a monotonically increasing greyscale on black and white devices.

To sum up, we required three-color mappings to cover the 11.6 bit scale gap: ACI-ENT-EVN at the smallest scales, BGN-POW-EVN at the mid-scales and the BGN-POW at the large scales (Table 2).

TABLE 2

Scale	Color mapping		
	Red	**Green**	**Blue**
Small	ACI	ENT	EVN
Mid	BGN	POW	
Large		**Cube-helix color-scale** POW	

C. Continuity between Scale Steps

The third constraint was our desire to preserve color continuity between scales steps in order to reduce user perplexity. Given the above two constraints, we adopted two techniques to preserving color-map continuity. The first was to assign the same index to the same color over two scale zones: EVN was assigned to blue over the small and mid-scales and BGN was assigned to red over the high-resolution scales (Table 2).

The second continuity technique was to achieve a smooth gradation in visual representation over the small and mid-scales by a graded combination of the ACI-ENT-EVN and BGN-POW-EVN color maps (see ratios in the right column, Table 1). The same technique was not effective when trying to achieve a smooth color-map transition between the mid and large scale spectrograms. Instead, the large transition in color-map was offset by taking a small-scale step from 0.2 s/px to 0.16 s/px.

D. Computational and Memory Costs

Although we have presented results for a 24-hour recording, typically source recordings may be any duration from 2 to 24 hours. We therefore give computational costs for a 60-minute recording and assume linear extrapolation to longer recordings.

- Size of 60 min. recording in WAV format: 310 MB
- Number of 0.2s segments per spectrogram: 3600×5 = 18000
- Index calculations per spectrogram: = 18000×256 = 4.6×10^6
- Count of index calculations for 8 spectrograms = 36.8×10^6
- Total size of index spectrogram data (zipped csv): 413 MB
- Total size of frame spectrogram data (zipped csv): 360 MB
- Total size of spectrogram images: 105 MB
- Total size of all the data: 310+773+105 = 1188 MB.
- Ratio of recording size to image tiles = 310:103 ≈ 1:0.33
- Ratio of recording to spectrogram data = 310:773 ≈ 1:2.5

The compressed size of the spectrogram csv files has been used because this gives a more accurate indication of the true information content. The png image files are compressed by default. If it is not required to keep the intermediate data files, then the cost of storing the image files is about one-third the size of the original WAV recording file.

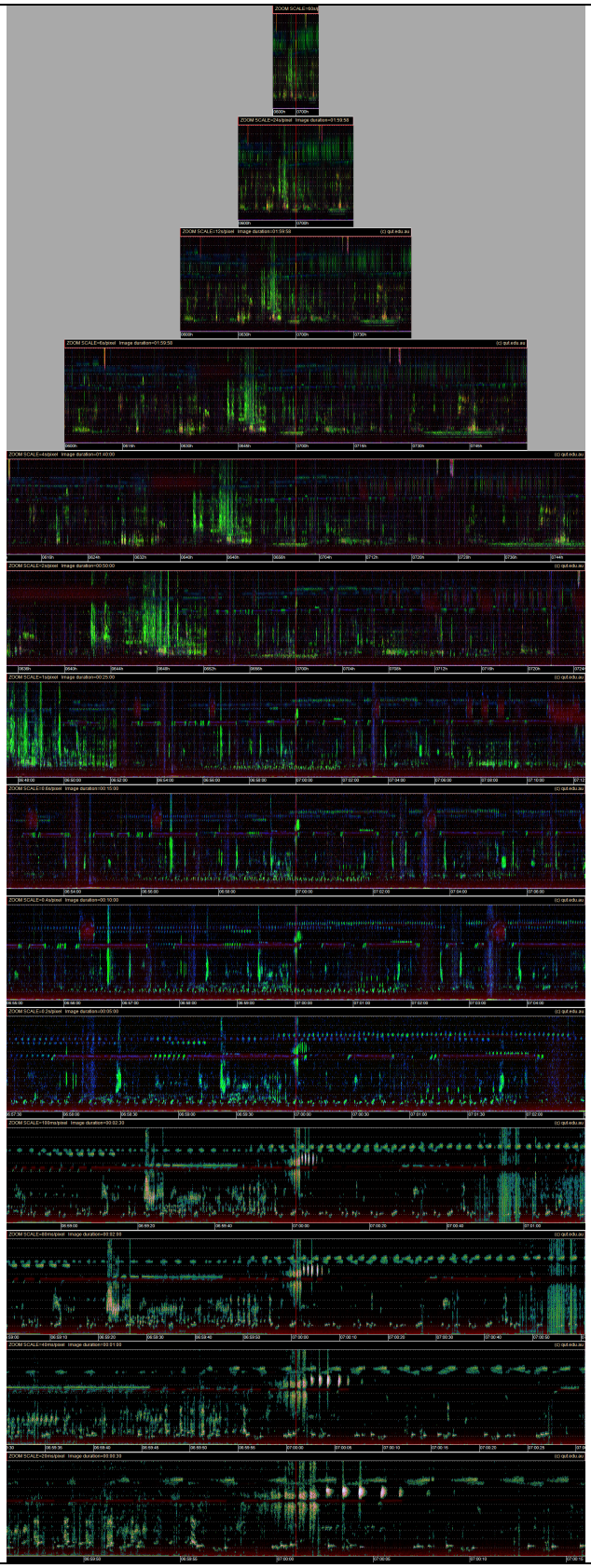

Fig. 4: A stack of zoomed images centered on 60 minutes into a two-hour recording.

VI. Case Studies

A. Sub-tropical rain forest recording

A complete zoom-stack of 14 images is shown in Figure 4. The purpose of this static image is to illustrate the scale steps (from 60 s/px to 0.02 s/px, a 3000× scale span) and the gradations in color-map that our investigation finally decided on. The recording was obtained in the Daintree National Park (north of Cairns, Australia), 6-8AM, November 2012. The stack is focused on 7AM (red vertical line). The image is better viewed at 200% zoom or more. The location is sub-tropical rain forest and the time-frequency soundscape is crowded with bird and insect vocalizations. For this reason, the structure of individual acoustic events does not become apparent until zoomed in to 2-4 s/px.

B. Koala recording

The recording in Figure 5 was obtained from St Bees National Park, St Bees Island, off the coast of Queensland, as part of a study to monitor the timing and frequency of koala bellows (*Phascolarctos cinereus*). We originally wrote a custom recognizer to detect koala bellows. However a text output of bellow 'hits' and times was not as informative as viewing the spectrograms at different scales to confirm context. It was possible to recognize a koala bellow at 60 s/px resolution because most calls have a 30-second duration. Consequently, one could scan a 24-hour recording in one screen-width image and zoom in to 1 s/px for confirmation. The 'magenta' event to the right of the top image is due to a helicopter. Note that it is quite distinct in appearance although the bandwidth and duration is similar to the koala bellow. Also visible is a cricket chorus at 3-4 kHz.

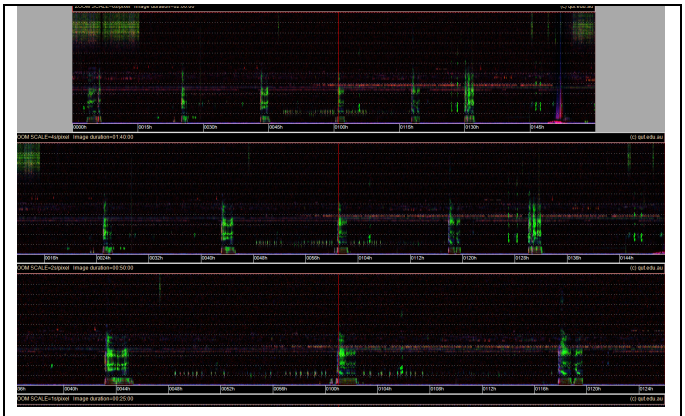

Fig. 5: Three image stack of Koala bellows **Top:** Six bellows in a two-hour recording. **Middle:** Five bellows in center 90 minutes of recording. **Bottom:** Three bellows in center 45 minutes of recording.

C. Kiwi recording

The recording in Figure 6 was obtained in a wildlife sanctuary near Wellington city (New Zealand) as part of a study of the little spotted kiwi (*Apteryx owenii*) [8]. The top image is scaled out twenty times compared to that of the bottom image but the distinctive kiwi call is still apparent in this nighttime recording. Apart from the louder (closer) call at 16 minutes, there is a quieter (more distant) call at 8 minutes which remains visible at the 5 s/px scale.

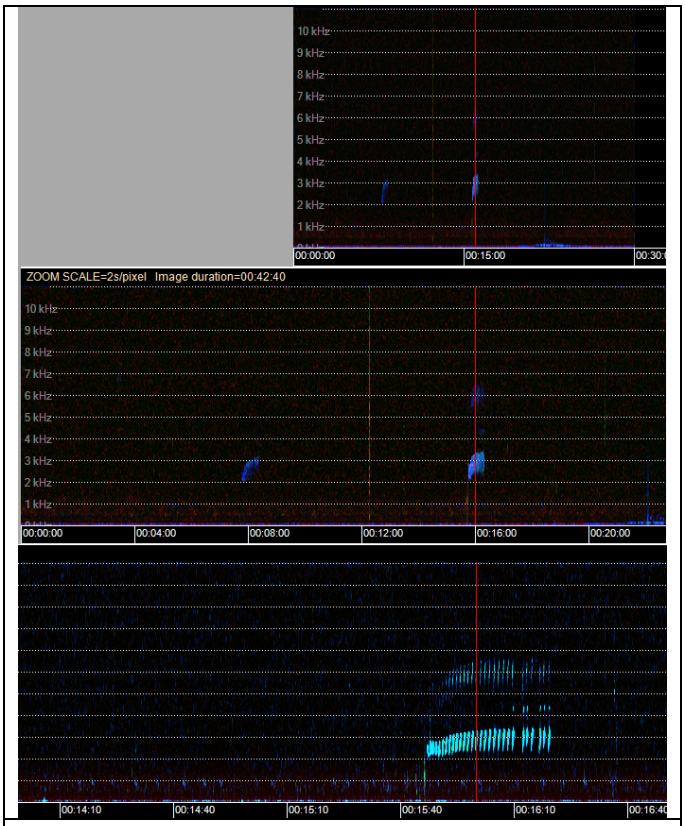

Fig. 6: Three spectrograms of the call of a little spotted kiwi. **Top:** Scale = 5 s/px. **Middle:** Scale = 2 s/px. **Bottom:** Scale = 0.25 s/px.

D. Wind and Artefacts

Figure 7 shows 8 hours of a recording from midday to 8PM at 60 s/px scale. Sustained high winds caused much clipping in this recording which would normally be excised for an ecological study. Clipping events are shown in red. Periods of wind are easily visible and can be quickly removed to leave just bird and insect activity.

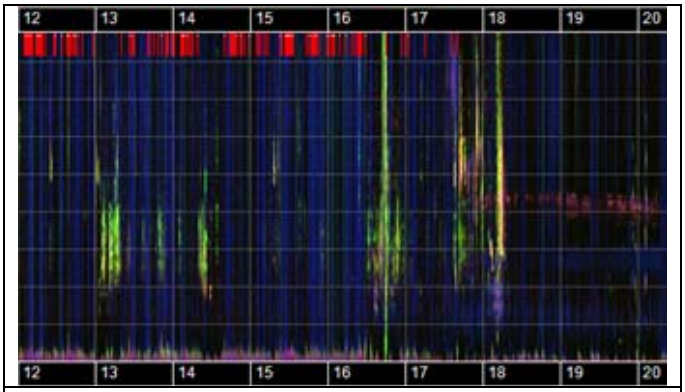

Fig. 7: Wind events (purple and blue) interspersed with bird calls (green) and insect calls (yellow-orange).

VII. Conclusion

We have described a technique to visualize and to navigate long-duration recordings of the environment. Navigation is achieved with the aid of a zooming facility, which treats the temporal dimension as a surrogate spatial dimension. Visualization is achieved by first extracting acoustic indices, statistics that reveal the distribution of acoustic energy in a way that is useful to ecologists. The principle advantage of spectrogram zooming is that it permits navigation and visualization at greatly different time-scales. This is important because the events that interest ecologists can range from a bird chirp lasting fractions of a second to the 24-hour diurnal soundscape. It is useful to have one interface that can display acoustic patterns on multiple scales. Spectrogram zooming invites integration with other systems of multiscale exploration of acoustic data [29, 30].

As noted in the introduction, we believe that, of the four approaches to visualization of long duration recordings (waveform, spectrogram, semantic icons, and chromagram), the spectrogram remains the most useful for ecological investigations because it makes time-frequency information explicit. However, even for environmental audio, it would be advantageous to combine all four approaches. For example, in our lab, we already combine long-duration spectrograms with parallel tracks indicating sunrise, sunset, and amplitude clipping (an indication of wind gusts). In fact, it is relatively easy to combine any kind of time-series data (temperature, rainfall, and salinity) with environmental spectrograms to assist interpretation of the audio.

We anticipate two areas of future work. First and most obvious is to explore additional indices, particularly indices which pick out characteristic acoustic patterns. For example, frog calls often have an amplitude-modulated oscillating structure and this feature can be extracted as an *oscillation index*. Second, is to integrate zooming spectrograms into a web-interface such as that described in [31]. However scaling up the availability of zooming spectrograms will require mass parallel processing [32].

Finally, it should be noted that the technique of producing false-color spectrograms has quite general application. In this paper, we have focused on the detection of birdcalls and therefore limited the temporal and spectral resolution accordingly. However, by appropriately adjusting the recording sample rate and window width, the method could be used to detect high frequency bird and insect calls and even the echo-locating calls of bats. Indeed the approach could be applied to audio from disciplines other than biology. Given an appropriate sampling rate, the critical steps are: 1. to select indices that capture acoustic information relevant to the discipline, and 2. to combine indices that capture, as far as possible, orthogonal acoustic information.

Acknowledgment

This research was conducted with the support of the QUT Samford Ecological Research Facility (SERF). The authors wish to thank Mark Cottman-Fields for IT support and stimulating discussion.

References

[1] A. Gasc, J. Sueur, S. Pavoine, R. Pellens, and P. Grandcolas, "Biodiversity Sampling Using a Global Acoustic Approach: Contrasting Sites with Microendemics in New Caledonia.," *PLoS ONE,* vol. 8(5), p. e65311, 2013.

[2] E. P. Kasten, S. H. Gage, J. Fox, and W. Joo, "The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology.," *Ecological Informatics,* vol. 12, pp. 50-67, 2012.

[3] W. Hu, N. Bulusu, C. T. Chou, S. Jha, A. Taylor, and V. N. Tran, "Design and evaluation of a hybrid sensor network for cane toad monitoring," *ACM Transactions on Sensor Networks (TOSN),* vol. 5, 2009.

[4] A. Farina, *Soundscape Ecology. Principles, Patterns, Methods and Applications*: Springer, 2014.

[5] R. D. Gregory and A. v. Strien, "Wild Bird Indicators: Using Composite Population Trends of Birds as Measures of Environmental Health," *Ornithological Science,* vol. 9, pp. 3-22, 2010.

[6] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, "What is soundscape ecology? An introduction and overview of an emerging new science," *Landscape Ecology,* vol. 26, p. 1213, 2011.

[7] S. Parsons and M. Towsey, "Report on a workshop (8-11 May 2012) to investigate the current status of environmental bio-acoustic monitoring," QUT ePrints, http://eprints.qut.edu.au/66572/, Queensland University of Technology, Brisbane 2014.

[8] A. Digby, M. Towsey, B. Bell, and P. Teal, "A practical comparison of manual, semi-automatic and automatic methods for acoustic monitoring," *Methods in Ecology and Evolution,* vol. 4, pp. 675–683, July 2013 2013.

[9] D. Stowell and M. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning.," *PeerJ,* vol. 2:e488, 2014.

[10] K.-H. Tauchert and K.-H. Frommolt, "Monitoring of booming bitterns (Botarus stellaris) by acoustic triangulation," *Bioacoustics,* vol. 21, p. 83, 2012.

[11] J. Wimmer, M. Towsey, P. Roe, and I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness," *Ecological Applications,* vol. 23, pp. 1419-1428, 2013.

[12] J. Sueur, S. Pavoine, O. Hamerlynck, and S. Duvail, "Rapid Acoustic Survey for Biodiversity Appraisal," *PLoS ONE,* vol. 3(12), p. e4065, 2008.

[13] N. Pieretti, A. Farina, and D. Morri, "A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index

(ACI)." *Ecological Indices,* vol. 11, pp. 868–873, 2011.

[14] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The Use of Acoustic Indices to Determine Avian Species Richness in Audio-recordings of the Environment," *Ecological Informatics,* http://dx.doi.org/10.1016/j.ecoinf.2013.11.007, 2013.

[15] T. Dang and N. Bulusu, "Lightweight Acoustic Classification for Cane-toad Monitoring," presented at the IEEE Proc. 42nd Asilomar Conf. on Signals, Systems and Computers, CA, USA, 2008.

[16] O. Kuhl and K. Jensen, "Retrieving and recreating musical form," presented at the Proceedings of the 4th International Symposium Computer Music Modeling and Retrieval, 2007.

[17] W.-Y. Chan, H. Qu, and W.-H. Mak, "Visualizing the semantic structure in classical music works," *IEEE Transactions on Visualisation and Computer Graphics (TVCG),* vol. 16, pp. 161-73, 2010.

[18] H. Kaper, S. Tipei, and E. Wiebel, "Data Sonification and Sound Visualization," *Computing in Science and Engineering,* vol. 1, pp. 48-58, 1999.

[19] H. Janicke, R. Borgo, J. S. D. Mason, and M. Chen, "SoundRiver: Semantically-rich Sound Illustration," *Eurographics,* vol. 29, 2010.

[20] J. Azar, H. A. Saleh, and M. A. Al-Alaoui, "Sound Visualization for the Hearing Impaired," *International Journal of Emerging Technologies in Learning,* vol. 2, pp. 1-7, 2007.

[21] M. Towsey, L. Zhang, M. Cottman-Fields, J. Wimmer, J. Zhang, and P. Roe, "Visualization of long-duration acoustic recordings of the environment," presented at the The International Conference on Computational Science (ICCS 2014), Cairns, Australia, 2014.

[22] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, and J. Sueur, "Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland.," *Ecological Indicators,* vol. 13, pp. 46–54, 2012.

[23] P. M. Fitts and J. R. Peterson, "Information capacity of discrete motor responses," *Journal of Experimental Psychology,* vol. 67, pp. 103-112, 1964.

[24] Y. Guiard, F. Bourgeois, D. Mottet, and M. Beaudouin-Lafon, "Beyond the 10-bit Barrier: Fitts' Law in Multi-Scale Electronic Worlds " presented at the Proceedings of IHM-HCI 2001, Lille, France, 2001.

[25] M. Towsey, "Noise removal from wave-forms and spectrograms derived from natural recordings of the environment," QUT ePrints, http://eprints.qut.edu.au/61399/, Queensland University of Technology, Brisbane 2013.

[26] DIMIN. *PanoJS3*. Available: http://www.dimin.net/software/panojs/

[27] D. A. Green, "A colour scheme for the display of astronomical intensity images," *Bulletin of the Astronomical Society of India,* vol. 39, p. 289, 2011.

[28] S. Silva, J. Madeira, and B. S. Santos, "There is More to Color Scales than Meets the Eye: A Review on the Use of Color in Visualization," presented at the IEETA/DETI, 11th International Conference Information Visualization (IEEE IV'07), University of Aveiro, Portugal, 2007.

[29] P. C. Wong and J. Thomas, "Visual analytics," in *IEEE Computer Graphics and Applications* vol. 24, 2004/9/1 ed, 2004, pp. 20-21.

[30] M. Sips, P. Kothur, A. Unger, H.-C. Hege, and D. Dransch, "A Visual Analytics Approach to Multiscale Exploration of Environmental Time Series," *IEEE Transactions on Visualization and Computer Graphics,* vol. 18(12), pp. 2899-2907, 2012.

[31] J. Zhang, K. Huang, M. Cottman-Fields, A. Truskinger, P. Roe, S. Duan*, et al.*, "Managing and Analysing Big Audio Data for Environmental Monitoring," presented at the 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE), 2013.

[32] A. Truskinger, M. Cottman-Fields, P. Eichinski, M. Towsey, and P. Roe, "Practical Analysis of Big Acoustic Sensor Data for Environmental Monitoring," presented at the 2014 IEEE International Conference on Big Data and Cloud Computing (BDCloud), Sydney, Australia, 2014.